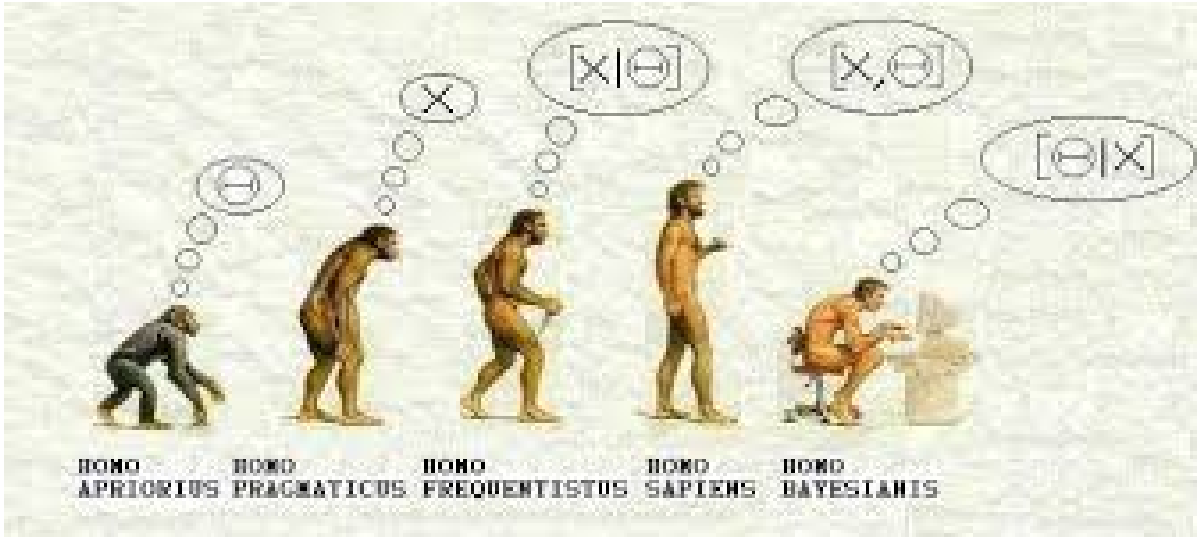
Bayesian Inference

Chapter 1: Bayesian basics

Conchi Ausín and Mike Wiper
Department of Statistics
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative Methods
Master in Mathematical Engineering

Objective



We introduce the different objective and subjective interpretations of probability and outline the main characteristics of the Bayesian approach in contrast with the usual frequentist methods.

Probability

- Probability theory developed from the work of Fermat and Pascal on games of chance in the 17th century.
- Mathematical, axiomatic theory developed by Kolmogorov (1933), but this is not interpretative.
- Different views of probability:
  - Classical approach
  - Frequentist approach
  - Bayesian approach
  - Logical probabilities, propensities: only for philosophers!

Classical probability

Classical probability stems from the work of Laplace (Bayes, ...) in the 18th century.

Principle of Insufficient Reason

If you have no prior reason to differentiate between the possible outcomes of an experiment, then you should assign the same probability to all of them.

Thus, probability is the number of favourable outcomes divided by the total possible number of outcomes.

Note that this is a subjective idea.

However, how do we do this in infinite or continuous sample spaces?

Frequentist probability

This approach was developed by Venn and von Mises at the end of the 19th century.

Given a repeatable experiment, the probability of an event is the limit of the proportion of times that the event will occur when the number of repetitions of the experiment tends to infinity.

What can we do in non-repeatable experiments?

Subjective probability

The formal ideas of a subjective definition of probability were devised in the 1920s by Ramsey.

A practical problem is that subjective probability is that we would like to be able to model human behaviour, but ...

humans are irrational and illogical and ...

not very good at probability!

Example

Federico is 35 years old, intelligent but not very imaginative and a bit boring. In college, he showed a lot of talent in maths but he wasnt very good at art.

Order the following statements about Federico in terms of their probability (1 = most probable, 8 = least probable).

1. Federico is a doctor and likes to play cards as a hobby.
2. He is an architect.
3. He is an accountant.
4. He plays a jazz instrument.
5. He reads Marca.
6. He likes mountaineering.
7. He is an accountant and plays a jazz instrument.
8. He is a journalist.

Rational agents

Therefore, the theory only considers *rational* agents.

- Axioms for rational behaviour:
  - Relative beliefs in the truths of propositions are transitive.
  - If we specify how much we believe A to be true, we implicitly specify how much we believe it to be false.
  - If we specify how much we believe A to be true and how much we believe B to be true assuming A is true, we implicitly specify how much we believe A and B to be true.
- Axioms imply that probability represents degrees of belief.
- This gives a general definition of probability. Different agents can have different probabilities for the same event ...

Bayes Theorem

... and when new data are observed, these beliefs can be updated using the following famous result:

Bayes Theorem

For two events A and B, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

More generally, if $\{A_i : i = 1, ..., k\}$ form a partition, then

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}.$$

The Monty Hall problem

Should you change doors?

Implicit assumption:

- the host always opens a different door from the door chosen by the player and always reveals a goat by this action because he knows where the car is hidden.

Solution using Bayes Theorem

Suppose without loss of generality that the player chooses door 1 and that the host opens door 2 to reveal a goat.

Let A (B, C) be the event that the prize is behind door 1, (2, 3).

$P(A) = P(B) = P(C) = \frac{1}{3}$.

$P(\text{opens } 2|A) = \frac{1}{2}$. $P(\text{opens } 2|B) = 0$, $P(\text{opens } 2|C) = 1$.

$$\begin{aligned} P(\text{opens } 2) &= P(\text{opens } 2|A)P(A) + P(\text{opens } 2|B)P(B) + P(\text{opens } 2|C)P(C) \\ &= \frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3} = \frac{1}{2} \end{aligned}$$

$$P(A|\text{opens } 2) = \frac{P(\text{opens } 2|A)P(A)}{P(\text{opens } 2)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

so $P(C|\text{opens } 2) = \frac{2}{3}$ and it is better to switch.

Statistical Inference

Various different approaches have been based on frequentist and subjective concepts of probability.

- Classical or frequentist inference
- Bayesian inference
- Likelihood based approaches and fiducial inference: not used these days

Classical Inference

- Developed by Fisher, Neyman and Pearson in the 1920s and 30s.
- Frequentist interpretation of probability.
- Inference is based on the likelihood $f(\mathbf{x}|\theta)$.
- θ is fixed. All uncertainty about $\mathbf{X}$ is quantified a priori.

- Inferential procedures based on asymptotic performance.
- Non prescriptive.
- Interpretation?
- What about *principles*?

Some principles

The sufficiency principle

If a sufficient statistic exists, then two samples of the same size with the same sufficient statistic provide the same information about θ.

Everyone in the world believes in the Sufficiency Principle.

The likelihood principle

In inference about θ, after the data **x** are observed, all relevant experimental information is contained in the likelihood function for the observed **x**.

This implies that two proportional likelihoods give the same information about θ.

The next principle is a consequence of the likelihood principle.

The stopping rule principle

In a sequential experiment, the evidence provided by the experiment about θ should not depend on the stopping rule.

The conditionality principle

Suppose that we have the possibility of carrying out two experiments E_1 and E_2 and that we select the experiment by tossing a coin. Then our inference for θ should only depend on the selected experiment.

Which of these principles seem plausible to you?

Example: a coin tossing experiment

You have a coin with $P(\text{head}) = \theta$ and wish to test $H_0 : \theta \leq 0.5$ vs $H_1 : \theta > 0.5$.

- You decide to toss the coin 12 times and observe 9 heads and 3 tails. Then the p-value is:

$$\sum_{i=9}^{12} P(i \text{ heads in 12 tosses} | \theta = 0.5) = 0.073$$

  and you don't reject the null hypothesis at a 5% level.

- You keep tossing the coin until you observe the third tail. This occurs on the 12th toss. Now:

$$p = \sum_{i=9}^{\infty} P(i \text{ heads before the 3rd tail} | \theta = 0.5) = 0.0325$$

  and you reject H_0.

In both cases, you have seen exactly 9 heads and 3 tails!

Bayesian inference: the prior distribution

Prior knowledge about θ is represented as a prior distribution.

What might be a reasonable prior for θ in the case of the coin tossing problem?

- $0 \leq \theta \leq 1$.
- Most coins are pretty symmetrical so maybe $f(\theta)$ should be symmetrical and centred on $\theta = 0.5$.
- One possibility is a beta prior distribution:

$$f(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$$

  for $a, b > 1$. This is symmetrical if $b = a$ and puts more weight around $1/2$ as a increases.
- How can we choose a? (later)
- What if we don't want a prior like this? (later)

Bayesian inference: updating

When data are observed, beliefs are updated via Bayes theorem

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)f(\theta) \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \end{aligned}$$

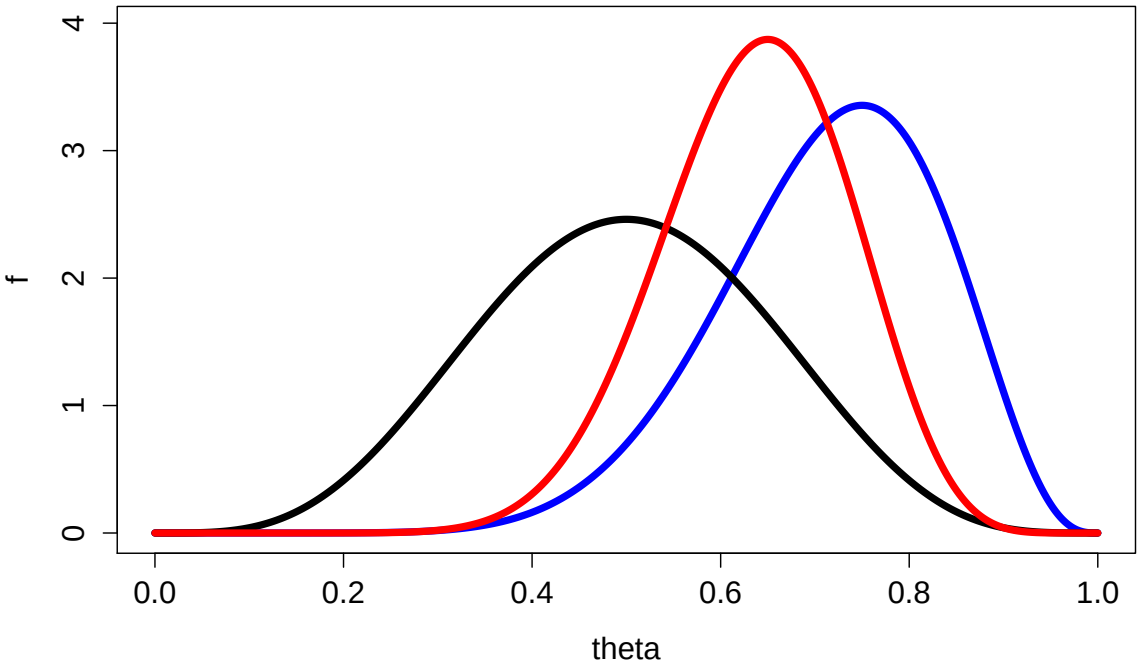
- Assume a symmetrical beta prior with $a = b = 5$.
- Then:

$$\begin{aligned} f(\theta|\mathbf{x}) &\propto \binom{12}{9} \theta^9(1-\theta)^3 \frac{1}{B(5,5)} \theta^{5-1}(1-\theta)^{5-1} \\ &\propto \theta^{14-1}(1-\theta)^{8-1} \end{aligned}$$

- What distribution is this?
- Another beta distribution. (that's lucky)
- Does the stopping rule make any difference?

Bayesian inference: the posterior as an average

The figure shows the prior (black), scaled likelihood (red) and posterior density (blue).



The posterior mean is

$$E[\theta|\mathbf{x}] = \frac{14}{22} = \frac{10}{22} \times \frac{1}{2} + \frac{12}{22} \times \frac{9}{12},$$

a weighted average of prior mean and MLE.

Bayesian inference: point and interval estimation

- For point estimates we could use the prior mean, median or mode.
- How can we justify these? (later)
- For interval estimates we can use a credible interval, i.e. an interval $[\underline{\theta}, \overline{\theta}]$ such that $P(\underline{\theta} < \theta < \overline{\theta}|\mathbf{x}) = 0.95$.
- The shortest such interval is a highest posterior density (hpd) interval.

- The MLE is $\hat{\theta} = 9/12 = 0.75$ and a 95% confidence interval (based on the standard normal approximation and on a binomial design) is $(0.505, 0.995)$.
- The mean of a beta(a,b) distribution is $\frac{a}{a+b}$ so the posterior mean is $\frac{14}{22} \approx 0.636$.
- A posterior 95% credible interval is $(0.430, 0.819)$.
- How do we interpret the two intervals?

Bayesian inference: prediction

Suppose that we wish to predict future observations, say $\mathbf{Y}$. Then

$$f(\mathbf{y}|\mathbf{x}) = \int f(\mathbf{y}|\mathbf{x},\theta) f(\theta|\mathbf{x})\, d\theta = \int f(\mathbf{y}|\theta) f(\theta|\mathbf{x})\, d\theta$$

in cases of conditionally i.i.d. variables.

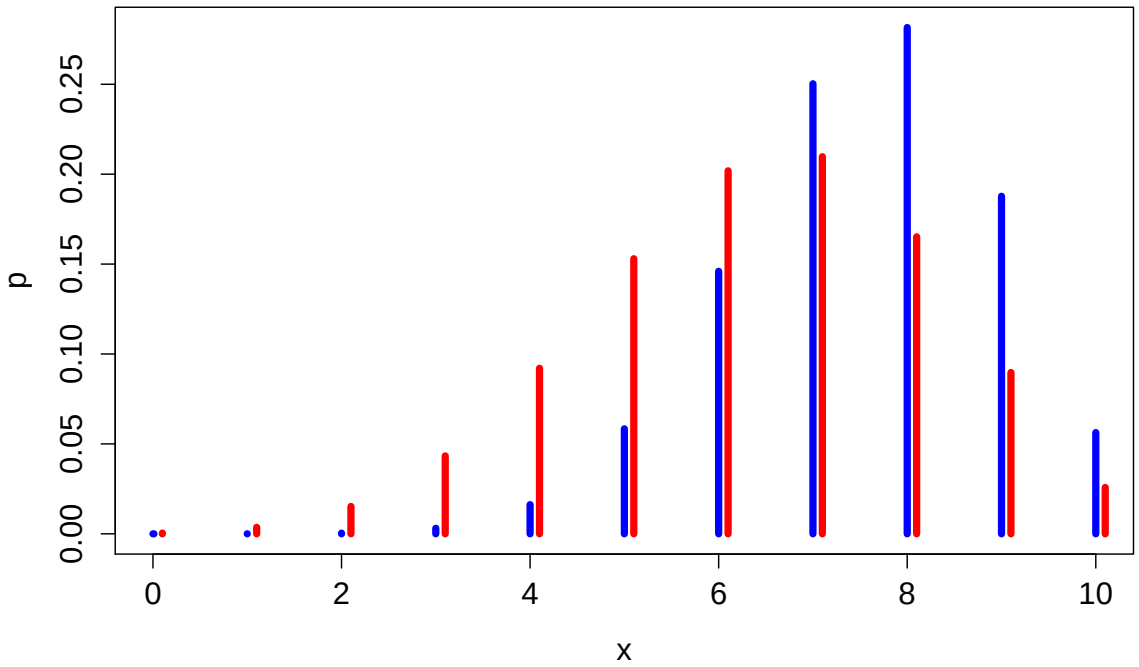
- Let's try to predict the number of heads, Y, in 10 further throws of the coin.
- We know that $Y|\theta \sim \text{Binomial}(10,\theta)$ independent of the results of the previous tosses.
-

$$\begin{aligned} P(Y=y|\mathbf{x}) &= \int_0^1 P(Y=y|\theta) f(\theta|\mathbf{x})\, d\theta \\ &= \dots \\ &= \frac{\binom{10}{y} B(14+y, 18-y)}{B(14,8)} \end{aligned}$$

for $y = 0, 1, ..., 10$.

Predictive distributions

The plot shows the classical "plug in" Binomial(10,0.75) predictive distribution and the Bayesian predictive distributions



Bayesian inference: hypothesis testing

- To test hypotheses H_0 and H_1, we define prior probabilities, $P(H_0)$ and $P(H_1)$, and associated parameter distributions $f(\theta_i|H_i)$ for $i = 0, 1$.
- Then calculate the posterior probabilities $P(H_i|\mathbf{x})$:

$$P(H_0|\mathbf{x}) = \frac{f(\mathbf{x}|H_0)P(H_0)}{f(\mathbf{x}|H_0)P(H_0) + f(\mathbf{x}|H_1)P(H_1)}$$

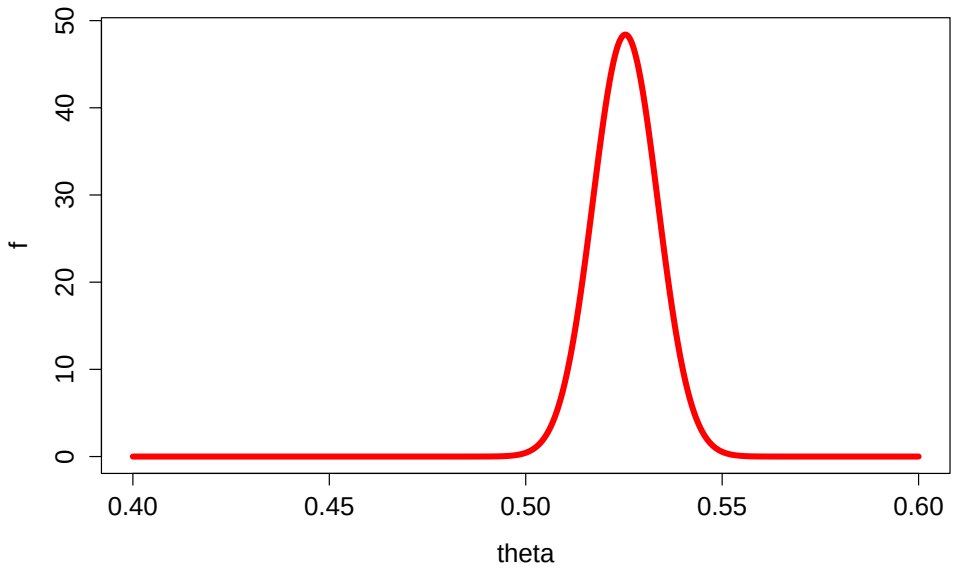
  where $f(\mathbf{x}|H_i) = \int f(\mathbf{x}|\theta_i, H_i)f(\theta_i|H_i)\, d\theta_i$ is the marginal likelihood under H_i.
- Now consider the type I and type II errors. Define a possible loss associated with each error.
- Choose the decision which minimizes expected loss.
- In practice this procedure may be difficult. (see later)

In our case, we have an implicit $P(H_0) = P(H_1) = 0.5$ with truncated beta priors for θ_0 and θ_1. Then $P(H_0|\mathbf{x}) \approx 0.095$ and $P(H_1|\mathbf{x}) \approx 0.905$.

What happens when I observe lots of data?

In a study on gender imbalance in China, Feldman et al (2008) reported 1923 male births and 1737 female births in Heilongjiang province of China.

- Let $\theta = P(\text{male birth})$. Then, the classical MLE is $\hat{\theta} \approx 0.5254$ and a 95% c.i. is $(0.5092, 0.5416)$.
- Assume a beta(5,5) prior for θ. Then the posterior mean is $E[\theta|\mathbf{x}] = 0.5253$ and a 95% credible interval is $(0.5091, 0.5415)$.



What about sensitivity analysis?

Typically, in a Bayesian analysis, we elicit prior information from (non statistical) experts.

- These experts will not be able to provide a full prior distribution, but maybe just a few quantiles ...
- Many priors will conform to this prior specification.
- Thus, we should examine robustness or sensitivity of our conclusions to the prior.
- If our conclusions vary a lot, we have to be very careful in prior specification.

What happens when I use a different prior?

- When we used Bayes theorem, we forgot about the denominator, that is the marginal likelihood,

$$f(\mathbf{x}) = \int f(\mathbf{x}|\theta) f(\theta)\, d\theta.$$

- We need to calculate this in order to derive the exact density, moments, ...
- Usually, this is not easy!

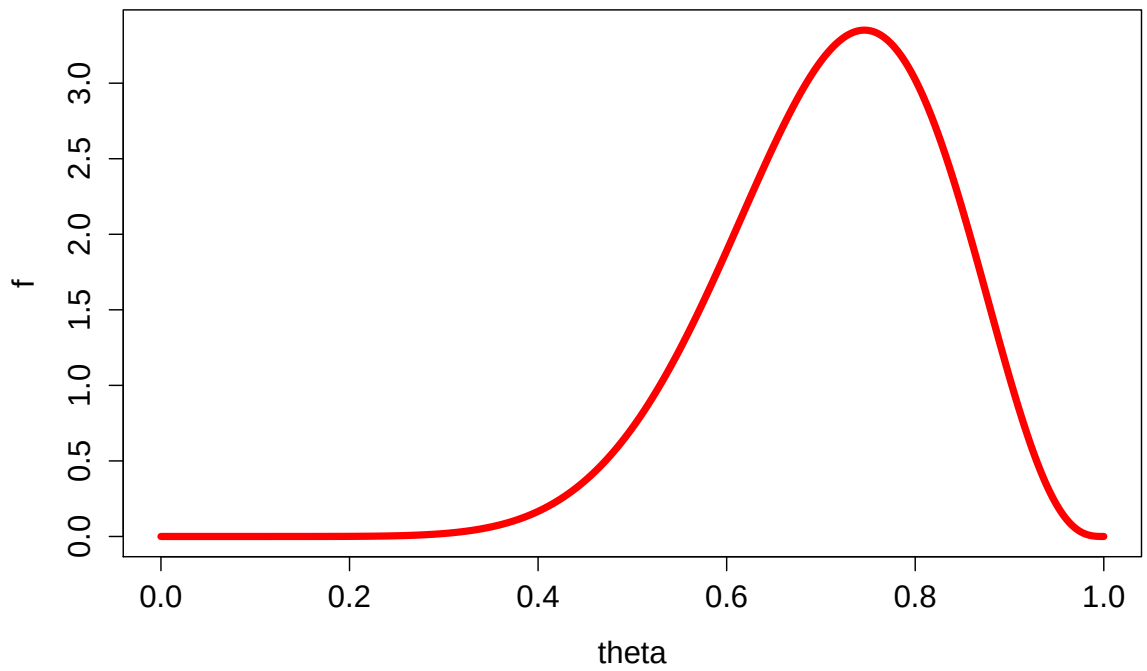
- Possible solutions are numerical integration or sampling based (Monte Carlo) methods.

Coin tossing again

- Assume that we use a truncated normal prior, $\theta \sim TN(0.5, 1)$.
- This has the same prior mean as the previous, beta prior, but more variance.
- Then, the posterior distribution is

$$f(\theta|\mathbf{x}) \propto \theta^9(1-\theta)^3 \exp\left(-\frac{1}{2}\left[\theta - 0.5\right]^2\right)$$

  and it is not possible to integrate this analytically to find the exact posterior density.
- The density and mean are relatively straightforward to calculate using numerical integration.



The estimated mean is 0.7117. Note that this is somewhat different to the beta case (0.6364).

Note however that if we want to calculate the variance, other moments, predictive distributions etc. we have to do further numerical integrations.

Monte Carlo

Assume that θ follows a distribution f. Then, if we can generate a sample from f, say $\theta_1, ..., \theta_N$:

- For a functional $g(\theta)$, we can estimate $E[g(\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} g(\theta_i) = \bar{g}$ say.
- If the variance $V[g(\theta)]$ exists, then we can estimate it as $V[g(\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} (g(\theta_i) - \bar{g})^2$.
- Remembering that sample averages are approximately normal (for large enough sample sizes) we can get a 95% c.i. for the true expectation as

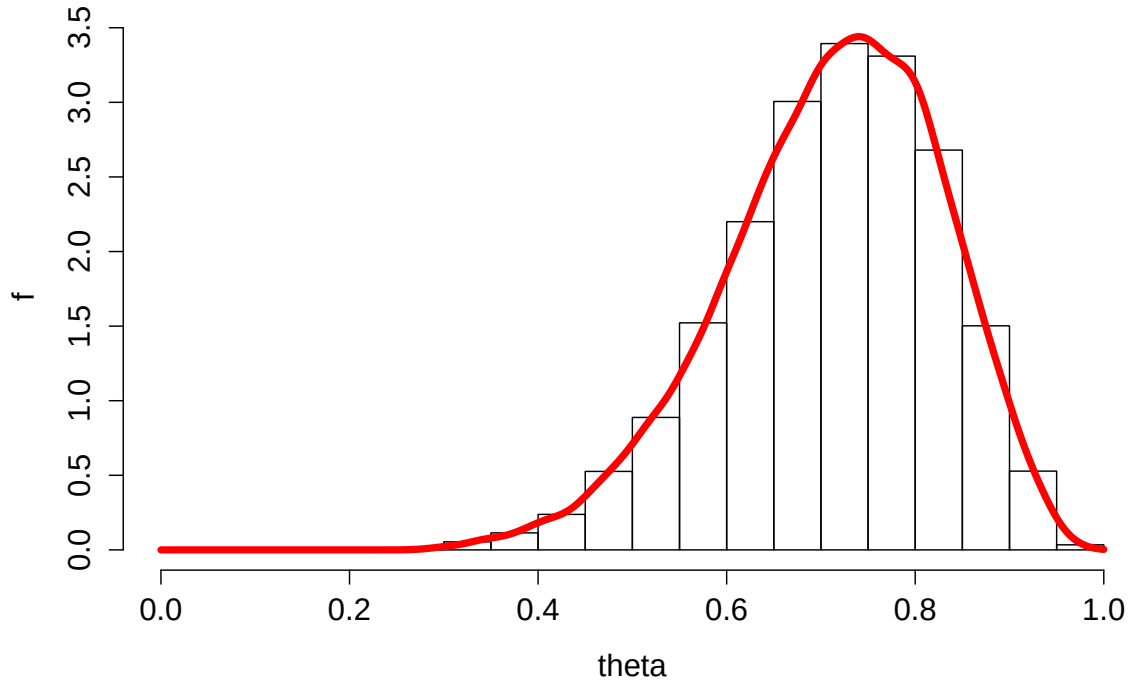
$$\bar{g} \pm 2 \text{ s.e.}(\bar{g}).$$

- We can estimate the cdf $F(\theta)$ using the empirical cdf of the sampled data and the pdf using a smoothed histogram (kernel density approximation).

Various methods are available for creating an (approximate) MC sample. (see later).

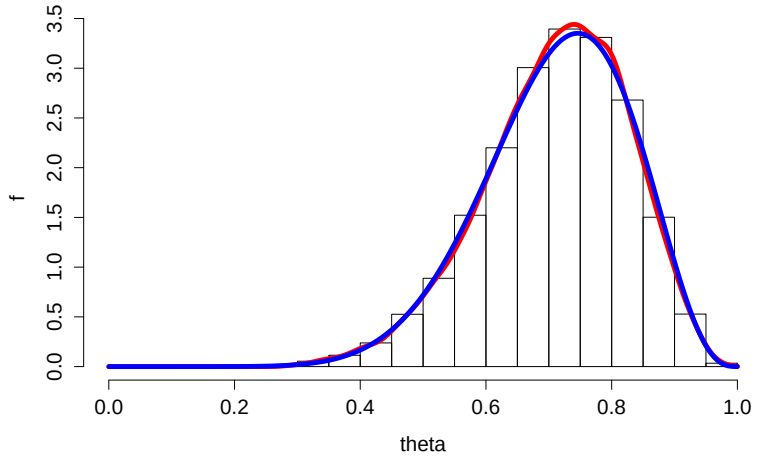
Coin tossing again

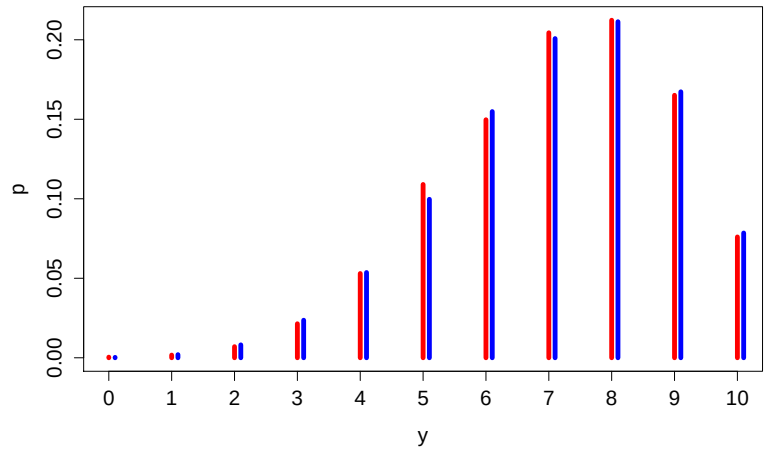
The following density was estimated using an (approximate) MC sample of 10000 data. The sample mean was equal to 0.7114.



Coin tossing again

It is very hard to appreciate the difference between simulated and numerically integrated results.





Summary and next chapter

We have seen an outline of the basic ideas behind Bayesian statistics and illustrated some of the ideas with a coin tossing example.

- We have seen that a beta(5,5) prior implied a beta posterior.
- This is a nice property as we know about beta distributions.
- When we opted for another prior we had to use numerical approaches.
- Next class: when do these nice priors exist?